SHAWN NEWSAM
Department of Electrical Engineering and Computer Science
University of California, Merced
Email: snewsam@ucmerced.edu

Shawn Newsam is an interdisciplinary researcher who works at the intersection of computer science and spatial data analysis. Most of his research has been centered around "pixels and location." He develops novel image processing, computer vision, and machine learning techniques for the automated understanding of georeferenced imagery. This work is motivated by the simple paradigm that if you know the geographic location of an image and you know what is in the image then you know what is on the surface of the Earth. This is, of course, the motivation behind the traditional field of remote sensing, which analyzes overhead images from satellites, aircraft, or, more recently, drones. However, he has also innovatively applied this paradigm to georeferenced ground-level images, a field which he has coined *proximate sensing*. He is currently interested in *spatially informed deep learning.*

Newsam is currently an Associate Professor of Electrical Engineering and Computer Science and Founding Faculty at UC Merced. He received a BS in EECS from UC Berkeley, an MS in Electrical and Computer Engineering from UC Davis, and his PhD also in ECE from UC Santa Barbara. Prior to joining UC Merced, he was a post-doctoral researcher with the Sapphire Scientific Data Mining group in the Center for Applied Scientific Computing at Lawrence Livermore National Laboratory. He is the recipient of a U.S. Department of Energy Early Career Scientist and Engineer Award, a U.S. National Science Foundation Faculty Early Career Development (CAREER) Award, and a U.S. Office of Science and Technology Policy Presidential Early Career Award for Scientists and Engineers (PECASE).

As a Founding Faculty at UC Merced, he has actively promoted spatial data analysis research and curriculum. He co-founded the UC Merced Spatial Analysis and Research Center (SpARC), which serves as the hub for spatial science research, analysis, education, visualization, spatial data archiving, and access to spatial science software and equipment for UC Merced and its partners.

He has been actively involved in the Association for Computing Machinery (ACM) Special Interest Group on Spatial Information and Analysis (SIGSPATIAL). ACM SIGSPATIAL is the premiere international community for computer scientists working with spatial information. He has served as the elected Vice-Chair (2014–2017) of the organization as well as the General (2016–2017) and Program Committee (2019–2020) Co-Chair of its flagship conference.

Setting the Spatial Data Science Agenda

In this position paper, I separately consider research and curriculum topic points related to spatial data science. For research, I describe my work and interest in spatially informed machine learning, particularly deep learning. For curriculum, I examine the role spatial data analysis can play in data science programs.

Research

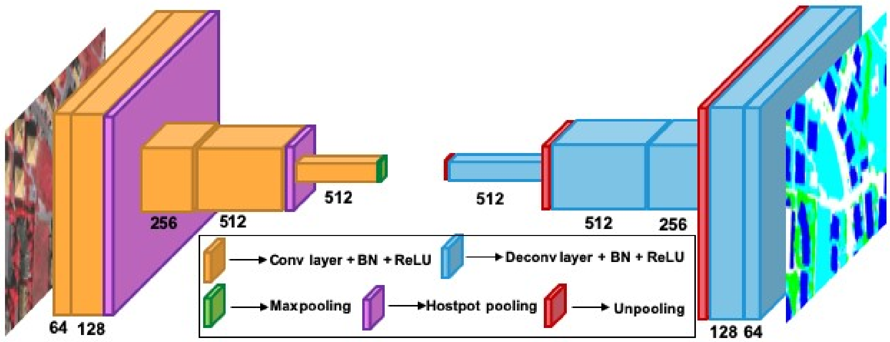
Enabled by large datasets and increased computational power, deep learning is transforming domains within computer science such as computer vision as well as beyond such as in the physical and environmental sciences. However, most deep learning models are purely data-driven and do not take into account knowledge about the data or processes that generate it. It is with particular interest that I read the comment in the symposium call about how recent work “has shown that spatially-explicit machine learning methods substantially outperform more general data [approaches?] when applied to spatial data even though this spatial component may seem of secondary importance at first glance.” As a computer scientist who develops novel machine learning algorithms for spatial data, I try to incorporate knowledge of the data or processes into my models. However, this is a particularly difficult challenge in the era of deep learning due to the black-box nature of the models. It is not obvious how to incorporate domain knowledge into the models. Spatially informed deep learning is currently a major focus of my research group at UC Merced.

Prior to deep learning, the state-of-the-art in computer vision for automated image understanding was based on local invariant features. Instead of analyzing the image as a whole, only the salient regions were considered. Image features (descriptors) were extracted from these regions and quantized so that the image could be represented as a bag of visual words (BOVW) similar to representing a text document as a bag of, say, English words. However, the spatial relationship between the visual words was lost. I therefore developed a technique termed spatial pyramid co-occurrence which characterizes both the absolute and relative spatial arrangements of the visual words[1,2]. The proposed model captures the spatial concepts of scale and distance. I showed the model not only outperformed the state-of-the-art in land use classification in overhead imagery but also improved performance on general computer vision problems such as object and scene classification.

More recently, I have investigated how spatial domain knowledge can be incorporated into deep learning models. One approach is to use location as a key to fuse different data sources for input to a model. My group has used this framework to combine aerial imagery with road maps to improve the semantic segmentation of the imagery (assigning a semantic class to each pixel). This can of course be extended to combine all sorts of spatial data. Another approach is to modify the loss function during model training to penalize, for example, the adjacent assignment of land cover/use classes that rarely appear next to each other.

The approaches above, though, do not make fundamental changes to the model architecture. This is the current focus of my most advanced PhD student. We recently developed a geospatially guided neural network (GeoNet) for improved overhead image segmentation. In this model, instead of performing max pooling in order to downsample the feature maps during the encoding stage of an encoder-decoder model, we perform hotspot analysis using the Getis-Ord Gi* statistic to decide which values to propagate. The motivation here is that, when downsampling, the most informative values should be propagated. The figure below shows the hotspot pooling layers in purple in a

GeoNet that takes as input an overhead image and outputs a semantic segmentation of this image into five classes: impervious surface, buildings, low vegetation, trees, and cars.



Our proposed GeoNet is shown to provide improved generalization over a traditional model. Generalization here means the performance of a model trained on ground truth data from one location when applied to another. Generalization is very important for scaling the analysis of geospatial data and is a fundamental challenge in data-driven machine learning. We are currently investigating other ways to incorporate spatial domain knowledge or analysis techniques into the model architectures. Moving forward, this is critical to realizing the full potential of deep learning for spatial data science.

Deep learning will remain a key component of data science at least in the near future and so there are additional challenges which must be addressed. Deep learning models, due to their datadriven and black-box nature, are typically difficult to interpret or explain and do not naturally provide measures of uncertainty. There is a rich history of these concepts in the spatial data analysis community and so, moving forward, there are great opportunities for synergy between the machine learning and GIScience communities in the context of spatial data science.

Curriculum

Data science programs in the form of undergraduate and graduate degrees, certificated, minors, etc. are appearing at educational institutions across the US. A discussion is needed on the role spatial data science can play in these programs. The 2018 Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report[3] by the National Academies of Sciences, Engineering, and Medicine does not identify spatial data analysis or thinking as key concepts for data science training but instead implicitly considers them as domain-specific considerations. But, should this be the case? Is there something fundamental yet unique about spatial data and thinking that require they be more central components of data science curriculum? The GIScience community needs to discuss and put forward a position on this.

References

[1] **Y. Yang and S. Newsam**, "Bag-of-visual-words and spatial extensions for land-use classification," *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS)*, 2010.

[2] **Y. Yang and S. Newsam**, "Spatial pyramid co-occurrence for image classification," *IEEE International Conference on Computer Vision,* 2011.

[3] **National Academies of Sciences, Engineering, and Medicine**, "Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report" The National Academies Press, 2018. https://doi.org/10.17226/24886.